Zoekmachines op internet

Het web was veel te

In 1994 zag zoekmachine Lycos het levenslicht, en pronkte met wel 50.000 webdocumenten. Google heeft er vandaag meer dan 2 miljard in zijn databanken zitten… Is het internet nu volledig in kaart gebracht, en hoe vind je het kaf tussen al dat koren?

Zoekmachines zijn immens populair: meer dan 85% van de surfers maakt er heel frequent gebruik van, in de hoop op die manier snel een geschikt document uit het web te kunnen opdiepen. In de praktijk blijkt echter een haast even groot percentage surfers dikwijls gefrustreerd te zijn door de resultaten van deze zoekmachines.

De ideale hitlijst?



Snel iets uitvissen over computers… Kunnen 61 miljoen hits volstaan?

Vanwaar die frustratie? Ofwel is je zoekterm te breed en krijg je een ontiegelijk groot aantal hits uitgespuwd. Ofwel tracht je het zoekterrein verder af te bakenen door meer gecombineerde, specifieke trefwoorden in te bouwen, zoals 'Nederland and (fietsen or wandelen) and not verkoop'… om dan plots vast te stellen dat je zoekterm met moeite nog enige hits oplevert!

Hoe komt het nu toch dat die zoekmachines je zo moeilijk de ideale hitlijst kunnen voorschotelen? Dat is aan twee oorzaken te wijten. Enerzijds ligt het aan de manier waarop zoekmachines hun gegevens op het web bij elkaar sprokkelen, in hun databanken stoppen en op vraag van de bezoeker weer te voorschijn halen. Anderzijds zit natuurlijk de massale hoeveelheid ongeordende informatie op het web er voor iets tussen, en die is een heel pak groter dan je wel zou vermoeden…

Gidsen

Laten we beginnen met de methodiek van zoekmachines: hoe worden de zoekresultaten beïnvloed door de werkwijze van de zoekmachines?

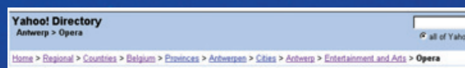
Web Site Directory - Sites organized by subject

Business & Economy B2B, Finance, Shopping, Jobs...	Regional Countries, Regions, US States...
Computers & Internet Internet, WWW, Software, Games...	Society & Culture People, Environment, Religion...
News & Media Newspapers, TV, Radio...	Education College and University, K-12...
Entertainment Movies, Humor, Music...	Arts & Humanities Photography, History, Literature...
Recreation & Sports Sports, Travel, Autos, Outdoors...	Science Animals, Astronomy, Engineering...
Health Diseases, Drugs, Fitness, Medicine...	Social Science Languages, Archaeology, Psychology...
Government Elections, Military, Law, Taxes...	Reference Phone Numbers, Dictionaries, Quotations...

Yahoo: moeder van alle webgidsen.

Strikt genomen moeten we een onderscheid maken tussen de webgidsen (of webdirectories) en de eigenlijke zoekrobots.

De eerste categorie is in hoofdzaak het werk van een team van redacteurs. Die hebben een wel erg originele job: dagelijks het web afspeuren naar zinvolle webdocumenten, en die vervolgens in een hiërarchische structuur trachten onder te brengen: van algemeen naar specifiek. Bezoekers van zo'n webgids kunnen dan in die hiërarchie afdalen tot wanneer ze bij een bruikbare pagina belanden.



De Antwerpse opera: een hele (Yahoo-)weg te gaan…

Ben je bijvoorbeeld op zoek naar 'voeding' en begin je bij de rubriek 'techniek' (stroombron), dan kom je op totaal andere webdocumenten uit dan wanneer je de rubriek 'gezondheid' (voedsel) aansnijdt…

Deze webgidsen hebben dan wel het voordeel dat ze met een meer uitgekiende selectie van webpagina's werken, ze hebben ook met een inherent nadeel af te rekenen. Het opnemen van webpagina's in hun databanken is immers een relatief langzaam proces, aangezien de medewerkers in principe eerst de inhoud van alle webpagina's nakijken. De databanken van webgidsen zijn dus per definitie erg beperkt!



Spinnen

De tweede categorie gaat geheel anders te werk. Zoekrobots als AltaVista en Google laten het speurwerk grotendeels over aan zogenaamde bots (ook wel spiders of crawlers genoemd). Dat zijn programma's die een bepaalde webpagina inlezen en die verder ook alle links volgen die op zo'n pagina vermeld staan. Ook de links van die nieuwe pagina's worden weer gevolgd, en zo gaat dat maar verder.

diep

Komt de bot terecht op een pagina die al in z'n databanken steekt, dan gaat hij na of die intussen is gewijzigd. Is dat zo, dan wordt de oude versie eruit gehaald en vervangen door de nieuwe. Zo'n bot gaat overigens met een ontstellende snelheid tewerk, zodat de databanken veel sneller aangroeien dan die van een webgids. Je kan dus wel veel meer hits verwachten, maar het is zeer de vraag of daar veel bruikbare adressen tussen zitten. In de praktijk komen echter meer en meer hybride vormen voor. ►

KAT-EN-MUIS

Search Engine Optimization Tips

Search engine optimization can be difficult and confusing. Knowing this we decided to put together some tips to help you with the process. They cover everything you need to know about optimizing your web pages for the search engines quickly and easily.

Search engine optimization tips listed in order of importance:

1. Potential site design/set up problems.
2. Selecting the correct keywords.
3. Your title tag.
4. Your copywriting.
5. Your meta tags.
6. Your images "alt" attribute.
7. What you should not do....
8. How long it takes to get listed.

Print All Tips

There's so much info in these tips you will probably find it easier to print them all out. We've placed them all on one long page formatted for easy printing!

Print all the tips!

Additional search engine optimization tips:

These tips are here because they can be useful to those that can implement them on their web site, but they are *not necessary* to achieve good listings in the search engines.

1. Hyperlinks.
2. Headings.

SEO-tips in overvloed op het web!

Vooral commerciële websites hebben er natuurlijk alle belang bij dat ze veel volk over de vloer krijgen: meer bezoekers betekent immers meer potentiële kopers. Een optimale rangschikking bij zoekmachines is de beste garantie om die bezoekers te krijgen. Daar worden zelfs speciale firma's voor ingehuurd, die de eigenaardigheden van de belangrijkste zoekmachines op hun duimpje kennen. SEP (search engine positioning) of SEO (search engine optimization) wordt zo'n service wel genoemd. Sommige nemen het daarbij niet al te nauw met de regels van de netiquette, en trachten vaak met slinkse middelen de zoekrobots te misleiden.

Een klassiek middeltje is natuurlijk het opnemen van misleidende trefwoorden en omschrijvingen in speciale codes van de webpagina. Maar soms worden ook heel aantrekkelijke woorden in witte tekst op een witte achtergrond onderaan de webpagina opgenomen, of in een minuscuul klein lettertype: niet zichtbaar voor mensen dus, maar wel voor zoekrobots.

Haast alle zoekrobots zijn intussen echter gewapend tegen zulke misleidingen en straffen die af met een slechtere rangschikking in de hitlijsten. Maar er zijn natuurlijk ook meer geavanceerde technieken… Cloaking (letterlijk: in een mantel hullen) is bijvoorbeeld erg populair. Sitebouwers optimaliseren eerst een bepaalde webpagina voor de belangrijkste zoekrobots, en plaatsen de verschillende versies op een webserver. Klopt op een bepaald moment een zoekrobot aan, dan merkt een programmaatje dat op die webserver draait dat op – bijvoorbeeld aan de hand van het IP-adres van de zoekrobot, en zal automatisch de webpagina worden opgediept die voor die bepaalde zoekrobot was geoptimaliseerd. Een menselijke bezoeker daarentegen krijgt de 'normale' webpagina te zien… Intussen zijn zoekrobots ook al bezig met dergelijke cloakingtechnieken te counteren! Het kat-en-muisspelletje kan dus nog wel even doorgaan…



ONZICHTBARE PAGINA'S ...

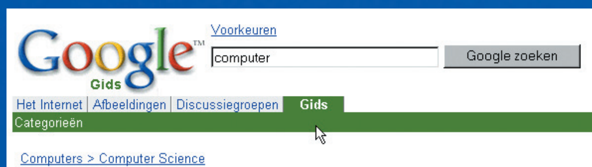
Zo worden de hitlijsten van Google samengesteld uit gegevens die door zijn bot werden aangebracht, maar tegelijk tapt Google ook dankbaar uit een ander vaatje: de databanken van een rasechte webgids, met name die van Open Directory [**http://dmoz.org**].

Van pagina naar hitlijst

De samenstelling van de hitlijsten bij de diverse zoekmachines wordt dus voor een groot deel bepaald door de hoeveelheid pagina's die ze kunnen bezoeken.

Maar dat is zeker niet de enige factor. Zo mogelijk nog belangrijker is hoe de zoekmachines die pagina's precies verwerken. Welke onderdelen van de webpagina's worden bijvoorbeeld ingelezen, en welke woorden uit die onderdelen worden vervolgens geïndexeerd en als opzoekbaar trefwoord in de databanken geplaatst? Zo hechten ongeveer alle zoekrobots veel belang aan paginatitels en aan de eerste paar regels tekst op de pagina, maar niet alle zoekrobots lezen bijvoorbeeld ook de tekst in die verschijnt als je met de muis over een afbeelding gaat (ALT-tags). Verder zijn er ook speciale codes op een webpagina die een gewone bezoeker niet te zien krijgt, maar die eventueel wel door een zoekrobot kunnen worden ingelezen. Zo kan de ontwerper van een webpagina zelf een eigen omschrijving (description) en trefwoorden (keywords) invullen die van toepassing zijn op z'n webpagina. De meeste zoekrobots houden ook rekening met deze speciale codes. Google is hierop een notoire uitzondering: heel wat sitebouwers misbruiken deze codes namelijk. Ze namen bijvoorbeeld heel populaire trefwoorden (als sex, mp3, enz...) op, in de hoop op die manier via de zoekrobots meer bezoekers naar hun pagina te lokken. In ons kaderstukje vind je trouwens nog een aantal andere slinkse trucjes waarmee men zoekmachines om de tuin – en naar eigen pagina's – tracht te leiden.

Twee elementen spelen dus al een belangrijke rol bij de samenstelling van hitlijsten: hoeveel en welke pagina's worden bezocht, en welke woorden van die pagina's worden geïndexeerd. Maar er is nog een derde hond in het kegelspel: hoe rangschikken zoekrobots de gevonden resultaten in de hitlijsten? Welke adressen komen bijvoorbeeld bovenaan de hitlijst te staan als je als trefwoord 'computer' opgeeft? Misschien wel de pagina waar dat woord het meest voorkwam? Of de pagina waarin dat woord in de url of in de paginatitel stond? Of misschien wel de pagina van degene die het meeste geld op tafel legde: 'gesponsorde links' wordt zo'n praktijk verbloemd omschreven, en steeds meer zoekmachines blijken hieraan toe te geven.



Google: zoekrobot én webgids tegelijk.

Sponsored Matches Info
Lowest Prices on computers at DealTime.com
Compare prices from thousands of stores and save instantly at DealTime! Find the best deal here.

Dell(TM) Computers on Sale Now
Order any Dell Home System on or before July 31st and you'll automatically be entered to win $50,000. No purchase necessary. Click for offer details.

Free Shipping this Weekend Only
Buy a select Gateway® desktop or notebook this weekend (Friday, Saturady, Sunday) and receive free shipping. For a limited time - offer expires 7/21.

Hot Deals on Computers at Tech Depot
Tech Depot by Office Depot lets you choose from over 60,000 low-priced **computer** and technology products, all with competitive shipping rates. Why wait? Buy your computers online now.

We found 48,708,951 results:
computer downloads ZDNet
Get downloads on computer

AltaVista: ook niet vies van sponsoring...

De aard van het (zoek)beestje

De ene zoekmachine is dus de andere niet: hun bots bezoeken niet alleen verschillende webpagina's, de indexering kan ook anders zijn en de algoritmen die de rangschikking van de resultaten in de hitlijsten aansturen, kunnen ook al verschillen! En alsof dat nog niet genoeg was, komt daar nog bij dat alle zoekmachines een aantal inherente tekortkomingen hebben!

Hoe snel de bots zich ook een weg kunnen banen door het web, ze kunnen nooit een up-to-date weergave zijn van alle webdocumenten! Het WWW is namelijk erg dynamisch: dagelijks komen er nieuwe pagina's bij en verdwijnen weer andere. Sommige zoekrobots trachten dit euvel min of meer op te lossen door dagelijks langs te lopen bij de webpagina's van grote en bekende nieuwssites (als die van CNN) en die opnieuw te indexeren.

Maar het grootste manco van alle zoekmachines is wel dat ze schromelijk onvolledig zijn, en dat brengt ons meteen tot de tweede grote factor die zoveel surfers vaak misnoegd naar de hitlijsten doet turen...

Het diepe web

Wie dacht dat kleppers als Google intussen zowat alle informatiebronnen op het web hebben bijgebeend, moeten we flink teleurstellen! Naar schatting zou Google intussen slechts 1/4 van de makkelijk toegankelijke webpagina's hebben verwerkt, en volgens sommige bronnen (zoals BrightPlanet) zouden de meer verborgen webpagina's – het zogenaamde diepe web - maar liefst 500 keer meer informatie bevatten. Anders gezegd: Google zou in dat ge-



Discussiegroepen: een gigantisch archief bij Google.

val niet meer dan 1 op 2.000 webdocumenten ontsloten hebben! Komt daar nog bij dat het web dagelijks met enkele miljoenen documenten aangroeit, zodat de kloof alleen maar groter dreigt te worden...

En dan hebben we het enkel nog maar over het web, want het internet is natuurlijk meer dan alleen het WWW. Vergeet niet dat er bijvoorbeeld ook nog tienduizenden verschillende nieuwsgroepen zijn, waarin dagelijks ontelbare berichten worden gepost. Heel wat van die berichten zijn intussen gelukkig wel gratis consulteerbaar: meer bepaald bij Google, via de knop **DISCUSSIEGROEPEN**.

Zoeken bij zoeksites

Metasearchers kondigen vaak met veel bombarie aan dat ze met deze tekortkoming komaf hebben gemaakt. In plaats van je zoekopdracht bij één enkele zoekmachine te concentreren, zorgen zij er wel voor dat je opdracht simultaan aan verschillende zoekmachines tegelijk wordt doorgegeven. De betere metasearchers halen meteen ook alle duplicaten uit de zoekresultaten en laten de gebruiker vervolgens toe eigen rangschikkingscriteria in te stellen. Een van de betere is ongetwijfeld Copernic 2001, waarvan je de basisversie gratis kan downloaden op [**www.copernic.com**].

In de praktijk blijken deze metasearchers met nauwelijks betere resultaten voor de dag te komen dan wanneer je een zoekrobot als Google apart aanspreekt.

Dat heeft vooral ook te maken met het feit dat de databanken van zoekmachines elkaar grotendeels overlappen. Als Google 25% van het zichtbare web heeft geïndexeerd, dan mag je daar hooguit 10 tot 15% bijtellen als je er de resultaten van alle bestaande zoekmachines bijneemt.

Search Progress - computer

Ah-ha.com	10
AltaVista	10
FAST Search (alltheweb...	10
FindWhat	10
HotBot	10
Lycos	0
Mamma.com	8
MSN Web Search	10
Netscape Netcenter	10
Open Directory Project	10
Yahoo!	10

Copernic 2001: verschillende zoekmachines simultaan aanspreken.

Het diepe web: hoe (on)zichtbaar?

Hoe komt het nu dat een ontelbare hoeveelheid webdocumenten zo goed als onbereikbaar blijken voor zoekmachines?

Om te beginnen bezoeken bots zelden webpagina's die meer dan drie niveaus diep verborgen zitten in een website. Ze bezoeken dan bijvoorbeeld wel de openingspagina, en pagina's waar die naar linkt, enz... maar na drie niveaus houden ze het dus meestal wel voor bekeken.



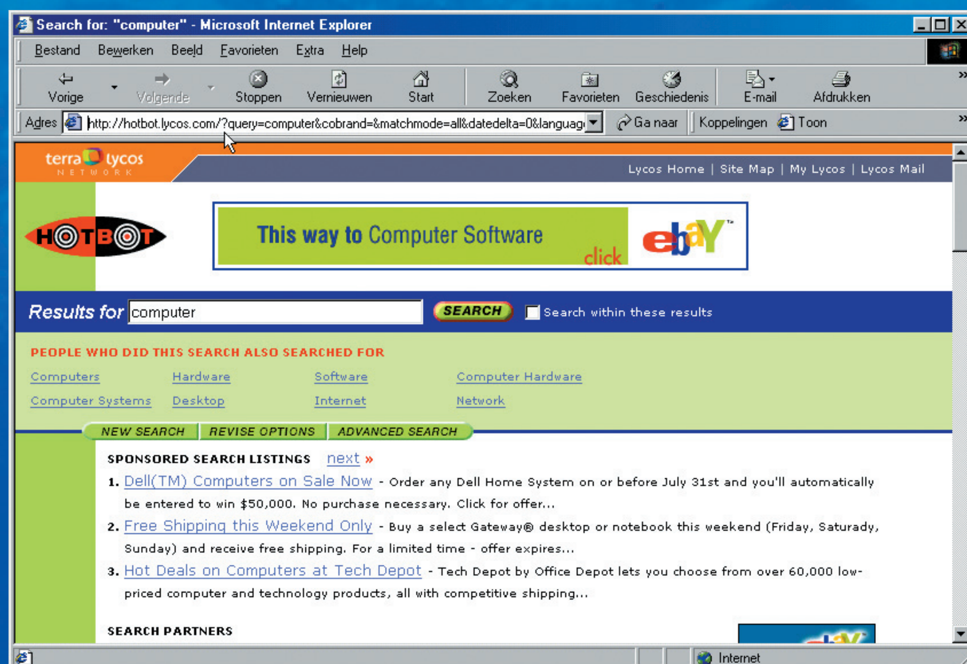
Geavanceerd zoeken bij Google: onvermoede mogelijkheden.

Verder wordt een klein deel van de webpagina's bewust geblokkeerd door sitebouwers: er is een wachtwoord nodig om toegang te krijgen, of ze hebben via een speciale code in hun webpagina's te kennen gegeven dat zoekbots bij voorkeur onmiddellijk rechtsomkeer maken. Neem je in de html-code van je webpagina bijvoorbeeld de volgende regel op, dan is dat voldoende om de meeste zoekrobots de toegang te ontzeggen en hen evenmin links van op die pagina te laten volgen: <meta name="robots" content="noindex,nofollow">.

Maar de toegang wordt niet altijd bewust geblokkeerd. Heel vaak slagen de zoekrobots zelf er niet in om webpagina's te ontsluiten. Een jaar of wat geleden bijvoorbeeld hadden heel wat zoekrobots de grootste moeite met webpagina's die met frames werkten. Een frame is een onderdeel van een grotere webpagina, dat op zich ook aparte schuifbalken kan hebben. Heel wat geframede pagina's geraakten om die reden dan ook niet in de databanken van die zoekmachines.

Tot voor kort bleven ook webdocumenten als doc's (Word), ppt's (PowerPoint), pdf's (Acrobat) en xls's (Excel) buiten beschouwing. Vooral Google heeft intussen ook werk gemaakt van de ontsluiting van deze – vaak heel interessante – documenten (zie **GEAVANCEERD ZOEKEN** bij Google).

Maar het zijn in eerste instantie de talloze dynamisch gegenereerde webpagina's die voor-



Dynamische, databankgestuurde webpagina's: lastige klanten!

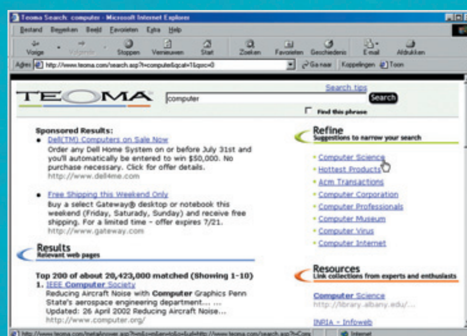
alsnog grotendeels buiten schot blijven! Dat zijn webpagina's die op het moment zelf worden samengesteld uit gegevens die zich in allerlei databanken kunnen bevinden. Het beste voorbeeld zijn overigens de pagina's met de hitlijsten van zoekmachines zelf. Het gros van het 'diepe web' moet je dus vooral in deze dynamische, databankgestuurde webpagina's gaan zoeken!

Optimalisatie zoekresultaten

Er is dus nog heel veel werk aan de winkel: enerzijds om de kwaliteit van de zoekresultaten bij te sturen, en anderzijds ook om dat immense, diepere web te ontsluiten!

De geavanceerde opties van zoekrobot AltaVista.

Elke zoekmachine die naam waardig biedt zijn bezoekers inmiddels ook geavanceerde zoekmogelijkheden aan. Dat gaat van Booleaanse operatoren (AND, OR, NOT en NEAR) over de mogelijkheid om sites op grond van domein of taal af te bakenen, tot 'familiefilters' die sites-met-rode-oortjes uit de hitlijsten trachten te weren. Maar de experimenten eindigen daar niet bij. Zo is AltaVista heel recent met een nieuwe service op de proppen gekomen: AltaVista Prisma. Zodra de bezoeker een zoekterm intikt, verschijnen er meestal twaalf extra, meer specifieke zoektermen die hem het terrein helpen afbakenen, zodat hij vlugger tot het gewenste resultaat kan komen.
Nog een stapje verder gaat zoekservice Teoma [**www.teoma.com**].



Teoma: eigen zoekcategorieën.

Die heeft talrijke zoekcategorieën in zijn databanken zitten, en als een zoekterm wordt ingetikt, gaat Teoma na of die niet in een of meer van zijn categorieën thuishoort. Is dat inderdaad zo, dan worden die categorieën aan de surfer aangeboden, en kan hij van daaruit zijn zoektocht verderzetten.

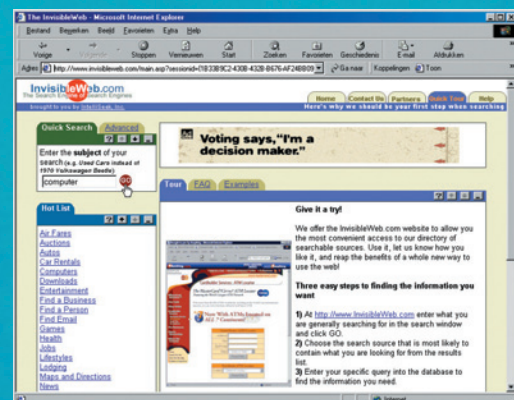


Vivisimo: zoekcategorieën on the fly!

Een gelijkaardig experiment tref je bij Vivisimo aan [**www.vivisimo.com**], maar hier worden de categorieën ter plekke door de zoekmachine zelf gegenereerd. Vivisimo sorteert de hits in een hiërarchische boomstructuur zodat je op een VERKENNER-achtige manier naar de gewenste resultaten kan bladeren. Lofwaardige pogingen, dat wel, maar af en toe zitten er ook serieuze missers tussen. Zo bijvoorbeeld genereerde Vivisimo bij het intikken van de zoekterm 'makreel' ook de categorie 'kant' omdat hij blijkbaar een aantal pagina's had aangetroffen waarin dat woord voorkwam (genre: 'makreel is zowel van de boot als vanaf de KANT te vangen').

Ontsluiting van het diepe web

Sommige experimenten trachten ook het grotendeels onontgonnen terrein van het diepe web te ontsluiten. Een van de eerste was InvisibleWeb.com [**www.invisibleweb.com**], dat naar eigen zeggen meer dan 10.000 databanken, archieven en zoekmachines ontsluit. Een heel pak daarvan zouden trouwens uit het diepe web komen.



The Invisible Web: toch niet zo onzichtbaar?

LexiBot van BrightPlanet is een recentere poging, en is intussen al aan versie 2.0 toe. Je kan het programma downloaden op [**www.lexibot.com**], en gedurende 30 dagen gratis uitproberen (daarna betaal je minimum $ 189). In eerste instantie dient het programma zich als een doordeweekse metasearcher aan, maar schijn bedriegt. Zo beperkt het zich niet tot de bekendste zoekmachines en zou het maar liefst zo'n 2.200 bronnen op het web aanboren. Bovendien wordt elk woord van de gevonden documenten op je lokale computer geïndexeerd, zodat je nadien makkelijk én snel filters of andere rangschikkingen in de hitlijst kan aanbrengen. Voor (grote) bedrijven is er zelfs een meer professionele oplossing: DQM (deep query manager): een on-lineservice die 35.000 gespecialiseerde databanken kan doorploegen, en waar meteen een even indrukwekkend prijskaartje aan hangt...
Het zal dus nog wel even duren voor de diepten van het web aan de oppervlakte komen, en zich gratis én efficiënt laten doorzoeken...

— Toon Van Daele —

AltaVista Prisma: pri(s)ma hulp?



LexiBot: tot in de diepten van het web?